

Shaping Up the Practice of Null Hypothesis Significance Testing

by Howard Wainer and Daniel H. Robinson

Recent criticisms of null hypothesis significance testing (NHST) have appeared in education and psychology research journals (e.g., Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989; Shaver, 1985; Sohn, 2000; Thompson, 1994, 1997; see also *Research in the Schools* [1998]). In this article we discuss these criticisms for both current use of NHST and plausible future use. We suggest that the historical use of such procedures was reasonable and that current users might spend time profitably reading some of Fisher's applied work. However, we also believe that modifications to NHST and to the interpretations of its outcomes might better suit the needs of modern science. Our primary conclusion is that NHST is most often useful as an adjunct to other results (e.g., effect sizes) rather than as a stand-alone result. We cite some examples, however, where NHST can be profitably used alone. Last, we find considerable experimental support for a less rigid attitude toward the interpretation of the probability yielded from such procedures.

In the almost 300 years since its introduction by John Arbuthnot (1710), null hypothesis significance testing (NHST) has become an important tool for working scientists. During the early

20th century, the founders of modern statistics (Sir Ronald Fisher, Jerzy Neyman, and Egon Pearson) showed how to apply this tool in widely varying circumstances, often in agriculture, and almost all very far afield from Dr. Arbuthnot's noble attempt to prove the existence of God. Cox (1977) termed Fisher's procedure "significance testing" to differentiate it from Neyman and Pearson's "hypothesis testing." He drew distinctions between the two ideas, but in our opinion, those distinctions are sufficiently fine that modern users lose little if we ignore them (see Huberty & Pike, 1999, for a discussion of the Fisher vs. Neyman-Pearson differences). The ability of statisticians to construct schemes that require human users to make distinctions that appear to be smaller than the threshold of comprehension for most humans is a theme we return to when we discuss α levels (the stochastic threshold commonly chosen for the dismissal of the idea of chance [Alberoni, 1962]).

With increasing use, both the strengths and limitations of NHST became more apparent. Given the breadth of the current discussion about the utility of NHST (see Harlow, Mulaik, & Steiger, 1997), it seems worthwhile to examine both the criticisms and the evidence and to provide a balanced, up-to-date summary of the situations for which NHST remains a viable tool as well as to describe those situations for which alternative procedures seem better suited. We conclude with some recommendations for improving the practice of NHST.

Most of the criticisms of NHST focus on its misuse by researchers and misinterpretation by research consumers rather than on inherent weaknesses. For example, a widely cited misinterpretation error concerns persons interpreting statistically significant results as being somehow equivalent to sci-

entifically significant results (Morrison & Henkel, 1970; Selvin, 1957). Researchers also badly misuse NHST when they incorrectly conclude that the null hypothesis of no difference or relationship is true when their results fail to reject it. We agree that any statistical procedure, including NHST, can be misused, but we have seen no evidence that NHST is misused any more often than any other procedure. For example, the most common statistical measure, the mean, is usually inappropriate when the underlying distribution contains outliers. This is an easy mistake; indeed such an error was made by Graunt (1662) and took more than 300 years to be uncovered (Zabell & Wainer, 2002).

The possibility of erroneous conclusions generated by the misuse of statistical procedures suggests several corrective alternatives. One Draconian extreme is to ban all such procedures from professional or amateur use. Another approach is to adopt the free marketer's strict *caveat emptor*. Both seem unnecessarily stringent, and it is hard to imagine any thinking person adopting either extreme: the former because it would essentially eliminate everything, the latter because some quality control over scientific discourse is essential. We favor a middle path—a mixed plan that includes both enlightened standards for journal editors as well as a program to educate users of statistical procedures. This article attempts to contribute to that education.

Some in the past (Schmidt, 1996) believed that the misuse of NHST was sufficiently widespread to justify its being banned from use within the journals of the American Psychological Association (APA). The APA formed a task force in 1996 to make recommendations about appropriate statistical practice. As a small part of its

The Research News and Comment section publishes commentary and analyses on trends, policies, utilization, and controversies in educational research. Like the articles and reviews in the Features and Book Review sections of *ER*, this material does not necessarily reflect the views of AERA nor is it endorsed by the organization.

deliberations it briefly considered this extreme proposal of banning NHST as well. However, banning NHST was never deemed to be a credible option.

Aristotle, in his *Metaphysics*, pointed out that we understand best those things that we see grow from their very beginnings. Thus, in our summary of both the misuses and proper uses of NHST we begin with the original intent of one of its earliest modern progenitors, Sir Ronald Fisher.

Fisher's Original Plan for NHST

Fisher understood science as a continuous and continuing process and viewed NHST in that context. He often used NHST to test the potential usefulness of agricultural innovations. He understood that science begins with small-scale studies with the purpose of discovering phenomena. Small-scale studies typically do not have the power to yield results of unquestioned significance. Moreover, Fisher recognized that the cost of getting rid of a false positive was small in comparison to the cost of missing something potentially useful. He knew that if someone incorrectly found that some innovation improved yields, others would quickly try to replicate it. If the results could not be replicated, the innovation would be dismissed.

This issue of replication led Fisher and Gossett (Student) to disagree over the character of experimental design. Fisher was as adamant in his demand for random sampling as he was for random assignment. However, Gossett was able to show that, for a single study, systematic sampling allows for a smaller error term, thereby increasing the precision of the parameter estimates. Fisher acknowledged that, in the short term, random sampling had its limitations. But Fisher's world of NHST was one where he always expected replications and, in that world, random sampling is better than systematic sampling because it provides a more accurate picture of the population's characteristics (see Student, 1938).

Fisher (1926) adopted a generous α of .05 to screen for potentially useful innovations "and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance" (p. 504). He understood that if a smaller α was used, say 0.001, then less dramatic improvements would be missed

and might not be rediscovered for a long time. Thus, .05 was used to screen for innovations that would then be replicated if found to be significant. Fisher (1929) went on to say,

In the investigation of living beings by biological methods, statistical tests of significance are essential. Their function is to prevent us being deceived by accidental occurrences, due not to causes we wish to study, or are trying to detect, but to a combination of many other circumstances which we cannot control. An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation. (p. 189)

There are two key parts of this quote—the mundane "once every twenty experiments" and the more important, "he knows how to design an experiment so that it will rarely fail to give a significant result." Fisher (1925, 1926, 1929) believed NHST made sense only in the context of a continuing series of experiments aimed at confirming the size and direction of the effects of specific treatments. Throughout Fisher's work he used statistical tests to come to one of three conclusions. When p (the probability of obtaining the sample results when the null hypothesis is true) is small (less than .05), he declared that an effect has been demonstrated. When it is large (p is greater than .2), he concluded that, if there is an effect, it is too small to be detected with an experiment this size. When it lies between these extremes, he discussed how to design the next experiment to estimate the effect better.

NHST as it is used today hardly resembles Fisher's original idea. Its critics worry

that researchers too commonly interpret results where $p > .05$ as evidence of no effect and rarely replicate results when $p < .05$ in a series of experiments designed to firmly establish the size and direction of the experimental effects. This is a science built largely of single-shot studies where researchers choose to reach conclusions based on these obviously arbitrary criteria. We should mention, however, that a strong counter-current to this practice is reflected in the Cochrane Collaboration (see www.cochrane.org), a database containing more than 250,000 random assignment medical experiments in which all of the included studies provide the information necessary for a meta-analysis. Such meta-analyses allow the formal concatenation of results, which then can yield more powerful inferences than would be possible from a single study. Robert Boruch, at the University of Pennsylvania, is currently organizing a parallel database for the social sciences; this effort is called the Campbell Collaboration (see www.campbell.gsie.upenn.edu).

We find it curious that NHST has been criticized for using arbitrary cutoff levels when at the same time the APA Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999) recommended that authors should report confidence intervals, which also use an admittedly arbitrary cutoff level of precision. The practice of basing scientific conclusions on single studies using arbitrary criteria, if widespread, could give NHST or any other method a bad name that could be avoided if researchers emulated Fisher's original plan. Nevertheless, there are additional ways in which NHST can be improved. We now examine how NHST has been misused and criticized unfairly and how it might be improved or used more appropriately.

Pointless Null Hypotheses

Schmidt (1996) and Thompson (1996) have complained that the typical null hypothesis is almost always false. We agree that NHST is being misused when it tests a null hypothesis in which the effect can go only in a single direction. Reporting a p value for a correlation computed for reliability and validity coefficients represents vacuous information (Abelson, 1997). If p values add nothing to the interpretation of results, leave them out, although sometimes a significant p value may just be scientific shorthand for a substantial effect

size. This occurs if our reaction to seeing a significant p value is to say to ourselves, “if the difference is statistically significant with that small a sample, it must be a huge effect.” Communicating effect size with p value and sample size is indirect, but sometimes such shorthand aids efficient communication.

Not all p values, however, are unimportant. Wainer (1999) mentioned several examples of research hypotheses where merely being able to reject the null would be a considerable contribution to science. For example, if physicists had been able to design an experiment that could reject the null hypothesis that the speed of light is equal in different reference frames that are moving at very different speeds, then a young Swiss patent clerk who suggested otherwise might have remained obscure.¹ Nevertheless, we agree that many of the null hypotheses tested in the research literature are false only in the *statistical* sense of the word, but as a practical matter could be treated as if they were true with little likelihood of any negative consequences. Newtonian physics is one example of a false hypothesis that under very broad conditions² could profitably be treated as true.

Usually, if large enough samples are obtained, p values can be made arbitrarily small. This criticism of NHST seems to us to be a valid one. If the only purpose of a hypothesis test is to canonize a small difference whose size and direction are of no interest, we agree that NHST is unnecessary. Furthermore, we generally agree with critics who suggest that it is the size and direction of observed differences that ought to be reported, and not only p values. We depart from complete agreement with such sentiments for those, admittedly rare, circumstances where such differences are of secondary importance and merely being able to reject the null hypothesis is the principal interest (e.g., H_0 : I am not pregnant).

We also depart from the critics in our belief that we ought to modify NHST to suit our modern understanding rather than to eliminate it. We discuss some plausible modifications in later sections.

The Role of Effect Sizes in NHST

An ordinal claim about the direction of the difference or relationship can be a substantial contribution (Frick, 1996). In some cases, however, knowing the direction of the effect is not sufficient to decide

whether an intervention is cost effective. In these situations, calculating the size of the effect (i.e., the degree to which sample results diverge from the null hypothesis, Cohen, 1994) is critical. Conducting NHST does not preclude the researcher from calculating effect sizes. Whereas NHST is useful in determining statistical significance, effect sizes are useful in determining practical importance. Of course, we would prefer to see all effect sizes accompanied with a confidence interval that indicates the precision with which that effect has been estimated. Nonetheless, we find the notion of choosing between either conducting NHST alone as is implied in most statistics textbooks or not conducting NHST and instead calculating effect sizes and confidence intervals as suggested by both Carver (1993) and Schmidt (1996) to be absurd. Both a frying pan and butter are useful on their own, but together they can do things that neither can do alone. So, too, it is with NHST, effect sizes, and confidence intervals. Researchers are free to use any statistical technique that sheds light on the interesting aspects of their data. Tukey (1969) recommended that “we ought to try to calculate what will help us most to understand our data, and their indications. We ought not to be bound by preconceived notions—or preconceived analyses” (p. 83).

Thompson (2003) reported that, over the past few years, 23 journals in education-related fields have instituted policies requiring authors to provide effect sizes in addition to p values. The reporting of effect sizes matches one recommendation of the APA Task Force on Statistical Inference that suggested that authors “always present effect sizes [and] add brief comments that place these effect sizes in a practical and theoretical context” (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). However, the most recent edition of the *APA Publication Manual* stops short of recommending that authors “always” present effect sizes because whether such a requirement is necessary is far from resolved. APA follows the policies of other successful institutions who understand not only that canonization requires that you be dead, but also that you must have been dead for a sufficiently long period of time (see Fidler, 2002, for a more critical examination of this decision).

Researchers should provide effect sizes or, for that matter, any other type of statis-

tical information that yields useful insights into the characteristics of data. However, requiring authors to always provide effect-size information may be overkill in those situations in which such information adds little to the correct interpretation of the data and, more dangerously, if it distracts or misleads readers. For example, a major use of NHST is in testing model fit, such as using a likelihood ratio to compare a restricted model to its more general parent. What does effect size mean in this context? Also, in some instances (e.g., medical research) it is a practical impossibility to obtain good estimates of effect size, because once a treatment is determined to be superior, researchers are ethically forbidden from using the inferior one. For example, the National Heart, Lung, and Blood Institute was forced to discontinue its 2-year study of propranolol in 1981 because it would have been unethical to keep the placebo group from receiving the treatment. By the way, the propranolol used in the study accounted for only $1/5$ of 1% of the variance, a seemingly small effect but with major medical importance (Rosenthal, 1991). We return to this issue of how some small effects have considerable practical significance later. For now, this circumstance provides a good illustration of two important ideas.

First, Will Rogers’s colorful caveat “What we don’t know won’t hurt us; it’s what we do know that ain’t” has important application in hypothesis testing. Indeed finding a significant, but inaccurate, direction of a difference or relationship has been called a Type III Error by Henry Kaiser (1970) in his Psychometric Society presidential address and was discussed many years earlier by Wald (1947). This suggests that accompanying an effect size by a suitably small p value is more than just an adornment.

The second issue worth mentioning is the question, “What is the effect whose size we are reporting?” In medical research one measure of a treatment’s effectiveness might be the number of people who do not get the disease who would have otherwise, or the number of people cured who would not have been: in short the causal effect of the treatment. Consider the ethical conundrum in trying to get a good estimate of the effect of a treatment. Obviously we want to know the direction of the effect of the treatment, and once we know it with

reasonable certainty, we are ethically bound not to use the inferior treatment. But how far can we continue with the experiment to be “sure enough”? Anscombe (1963) proposed a modification to the typical Neyman-Pearson formulation more in keeping with medical needs that forms a model for the type of flexible approach we support. Anscombe pointed out that we are not interested in the asymptotic probability of error; rather he observed that for any medical treatment there are a finite number of patients treated. A small number of them are treated as part of the clinical trial; the rest are given the treatment that the clinical trial decides is “best.” If we use too small a number of patients in the trial, the decision of which treatment is best is more likely to be in error. And if so, all the rest of the patients will be given the wrong treatment. If we use too many patients in the trial, then all the patients in the trial on the other treatments will have been mistreated unnecessarily. Anscombe proposed that one criterion of analysis, one “effect,” should be minimizing the total number of patients, both those in the trial and those treated afterwards, who are given the poorer treatment.

In the field of education, suppose we are interested in comparing two different curricula used in an elementary science class, where one leads to better and longer-lasting learning than the other. How many randomized field trials should be conducted before we determine the winner? In this case minimizing the number of students who receive the inferior curriculum is a worthy goal.

Finally, there are also some situations in basic rather than applied educational research where obtaining a large or practical effect is not necessary or useful. McKeachie (personal communication, March, 19, 1997) noted that effect sizes are mostly useful for

research that is directed toward decisions with some immediate practical consequences. As I see it, much research is concerned with developing or testing theory. If it is to test an existing theory, even a small difference should increase one’s confidence that the theory has some validity. Similarly if you are contributing to theory development, the size of the result is not so important as its heuristic value in stimulating thinking, which may then be tested by further research.

Thus, in some cases, researchers can or should only look for effect direction and not effect size.

There are even very practical situations in which effect size is known in advance to be very small and only direction is of interest. For example, consider an application of what Box and Wilson (1951) called “evolutionary variation in operations” (EVOP) in which slight variations in manufacturing procedures are tried and the direction of their effect noted—does it improve matters or make them worse? The variations are never large because the costs of a major screwup are too serious. If the direction of change is an improvement, then further changes of that sort are made. If things get worse, subsequent changes are made in another direction. As an example, suppose a manufacturer of paper is using EVOP and introduces experiments into the production run. The humidity, speed, sulphur, and temperature are modified slightly in various ways. The resulting change in paper strength cannot be great and still produce a salable product. Yet some of these slight modifications may yield a significant increase, which becomes then the stage for another experiment. The results of each stage in EVOP are compared to previous stages. Experiments with seemingly anomalous results are rerun. The experiments go on forever, for there is no final “correct” solution. This matches closely the scientific enterprise in which the sequence of experiments followed by examination and reexamination of data has no end. One noteworthy example in education intervention research is Robert Slavin’s *Success for All* program where numerous slight modifications have been made over a 15-year period (Slavin & Madden, 2001).

Finally, there may be circumstances in which requiring authors to provide effect-size information may be inappropriate. Robinson, Fouladi, Williams, and Bera (2002) had college students read research abstracts and found the inclusion of effect sizes led readers to overestimate the importance of research results. Including effect-size information may also cause readers to fail to distinguish between significant and nonsignificant effects for single-study conclusions. For example, suppose a small, spurious effect is reported with a confidence interval and the author goes on to discuss the size of the effect as if it were meaningful. This mismatch between the results of

statistical tests and researchers’ interpretations of them has been termed a Type IV Error (Marascuilo & Levin, 1970).

The educational and psychological research literature contains many examples of researchers providing effect sizes for nonsignificant effects. In perusing some of the most recent issues of such journals, we found several examples such as, “students showed only a slight and statistically nonsignificant mean increase in their performance: simple scoring, $t(13) = 0.98, p = 0.17$, descriptive effect size = 0.44” (Derry, Levin, Osana, Jones, & Peterson, 2000, p. 759). Derek Briggs (2001), in his large-scale study of effects of coaching on college admission exams, found an increase of 11% of a standard deviation on the ACT-English test (his Table 7) but in the text pointed out that it was not significant. But the same table contained an estimate of the effect of coaching on ACT-Reading of -11% (coaching was associated with lower ACT scores on reading). This effect is significant. A hurried reader might assume that two effects of the same size (albeit in different directions) on the same test battery are significant. We recommend that, if authors must publish single-shot studies, they follow a two-step procedure where first the likelihood of an effect (small p value) is established before discussing how impressive it is (Robinson & Levin, 1997), and that such discussions be made with great clarity. Of course, the danger of misinterpretation is reduced dramatically if authors instead publish multiple-shot studies.

How Do We Determine Practical Significance?

Determining the impressiveness of a particular effect size requires interpreting what the size of the effect really means. Royer (2000) cautioned, “[I]t is important to note that measures of effect sizes do not directly translate into indications of practical importance” (p. 239). Over 20 years ago, Glass, McGaw, and Smith (1981) similarly warned,

Above all else, this is clear about magnitudes of effect: *There is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as “small,” “moderate,” “large,” and the like.* Dissociated from a context of decision and comparative value, there is no inherent value to an effect size of 3.5 or .2. Depending on what benefits can be achieved at what cost, an effect size of 2.0

might be “poor” and one of .1 might be “good.” After decades of confusion, researchers are finally ceasing to speak of the regions of the correlation coefficient scale as low, medium, or high. The same error should not be repeated in the case of the effect-size metric. (p. 104)

Despite this warning, we continue to see researchers interpreting effect sizes according to arbitrary criteria, “[t]he effect size (.63) for this study falls between a medium and a large effect—a reasonable size for the behavioral sciences” (Orange, 1999, p. 34). Knapp and Sawilowsky (2001) recently stressed the importance of content expertise in explaining statistical findings. They noted that the logo of the Royal Statistical Society is “Allis Exterendum,” (“Let others thrash it out”). In other words, people with content expertise should be the ones who decide if results are practical and meaningful. Perhaps authors lacking such expertise should simply present their findings without commentary about impressiveness.

It is common in education to present an r^2 as an effect size, yet there are circumstances in which the r^2 is close to 0.00 and the practical significance is considerable. For example, the Steering Committee of the Physicians’ Health Study Research Group (1988) concluded that aspirin reduced the risk of heart attacks. The study involved 22,071 physicians; approximately half were given an aspirin every other day over a 5-year period, and physicians in the control group were given a placebo. The statistical results indicated heart attacks occurred more frequently in the placebo condition, $\chi^2(1, N = 22,071) = 25.01, p < .00001$. The r^2 was .001, indicating that less than 1% of the variance was accounted for. Some might be tempted to conclude that the results are practically nonsignificant because of the small effect size. However, viewing the same results using Rosenthal and Rubin’s (1982) binomial effect size display (see Table 1) shows that 85 fewer persons experienced a heart attack when they took aspirin every other day (Rosenthal, 1995; Rosnow & Rosenthal, 1989). In this case, at least 85 physicians would testify to the practical significance of an effect whose r^2 is zero to two decimal places. This illustrates the concept of a study’s clinical, rather than statistical or practical, significance (Thompson, 2002a).

Table 1
Binomial effect size display of $r^2 = .001$

Condition	No Heart Attack	Heart Attack	Total
Aspirin	10,933	104	11,037
Placebo	10,845	189	11,034

Another example of a small effect size and a potentially important finding comes from physics. Webb, Murphy, Flambaum, Dzuba, Barrow, Churchill et al. (2001) recently conducted a study and found that the fine structure constant that reflects how closely packed atomic particles can be is different than it used to be. They reported a p value of less than .05 and also an effect size: it has changed only one part in 100,000 over the past 12 billion years. However, Mario Livio, theorist at Space Telescope Science Institute in Baltimore, commented, “If confirmed, this is a sensational result. This is really probably one of the major breakthroughs we have seen.” (Kolata, 2001) Interpreting effect sizes is subject to the same misuse as interpreting p values. Our recommendation is that interpretation must be made only by those with the requisite subject-matter expertise.

Arbitrary α Levels

Nickerson (2000) mentioned the “arbitrariness of the decision criterion” as one of many criticisms of NHST (p. 269). We agree that researchers should not be bound by the chains of $\alpha = .05$. In a very real sense .05 is an anachronism. It was settled on when p values were hard to compute and so some specific values needed to be provided in tables. Now calculating exact p values is easy and so the investigator can report “ $p = .06$ ” and leave it to the reader to decide how significant it is. The fact that many misuse NHST by simply making reject or fail-to-reject decisions on single studies is probably due to the Neyman-Pearson legacy of such dichotomous decisions. We recommend that p values should be reported as Fisher suggested so that the readers can decide for themselves whether the direction of an effect can be established. However, if researchers attempt to interpret the p values in a dichotomous (Neyman-Pearson) way, such as “due to chance” or “not due to chance,” they should select an α level a priori and explain why it was chosen. The level of α chosen should correspond to the researcher’s threshold for the

dismissal of the idea of chance for that particular null hypothesis. A person’s threshold may certainly change given the stakes of the hypothesis that is tested. Fisher (1970, originally published in 1925) himself stated that

no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (p. 42)

Tukey (1969) discussed the potential problems of using different α levels for different contexts.

Need we—should we—stick to $p = .05$ if what we seek is a relatively pure list of appearances? No matter where our cutoff comes, we will not be sure of all appearances. Might it not be better to adjust the critical p moderately—say to .03 or .07—whenever such a less standard value seems to offer a greater fraction of presumably real appearances among those significant at the critical p ? We would then use different modifications for different sets of data. No one, to my knowledge, has set himself the twin problems of how to do this and how well doing this in a specific way performs. (p. 85)

If researchers are conducting small-scale studies to be included as part of a continuing series of studies, then using .05 as an α level seems appropriate as a screening device (Robinson, Funk, Halbur, & O’Ryan, in press). However, if researchers are conducting one-time studies that have serious consequences, much smaller α levels should certainly be used. But basing any high-stakes decisions on a single study should be an unusual circumstance indeed.

What if $p = .06$?

Referring to outcomes where $p < .05$ as significant and where $p > .05$ as nonsignificant is problematic when p values are close to .05. Journal article authors are split as to how to treat effects when p is small but not statistically significant ($.05 < p < .10$).

Sometimes researchers treat these effects as though they were significant: “[t]hese analyses indicated that cyberstudents ($M = 9.97$) had a more external locus of control than conventional students ($M = 8.17$), $F(1, 4) = 7.22, p < .06$ ” and “[c]umulative scores were correlated with hours per week studying with . . . the total hours per week studying for the course ($r = .42, p < .08$)” (Wang & Newlin, 2000, p. 140). Other researchers choose a “marginal” modifier: “[h]owever, there was a marginally significant interaction between grade and the experimental manipulation, $F(1, 176) = 3.02, p < .084$ ” (Rubman & Waters, 2000, p. 507). Still others treat these small but statistically nonsignificant effects as if they represented no effect at all: “[a] repeated measures ANOVA . . . revealed no interactions for teachers by treatment, $F(2, 8) = 3.56, p = .06$ ” (Johnston, 2000, p. 251), “a one-way ANOVA did not yield a main effect of group, $F(3, 268) = 2.41, p < .07$ ” (Schneider, Roth, & Ennemoser, 2000, p. 291), and “[t]here was no significant difference of overall performance between the students of the two groups, $F(1, 411) = 3.07, p = .08$ ” (Ni, 2001, p. 408).

As previously noted, Fisher used the .05 level as a heuristic because he knew that if a potentially useful treatment were discovered, someone would replicate it and show it to be useful. We believe that p values should be interpreted in the context of a series of experiments. If $p = .06$, then the researcher should ask if the effect is of potential interest to explore further. Fisher always attempted to improve the design when p values were between .05 and .2.

In quantitative research, consistent small probabilities from several studies that indicate effects in the same direction are needed to conclude the direction of an effect. Statistically significant results that are replicated provide the basis of scientific truth (Tukey, 1969). As for describing results from single studies, Tukey (1991, as cited in Abelson, 1995) proposed that we might use additional words besides significant or nonsignificant to describe our reluctance to bet on the direction of the true difference or relationship. For example, if p is greater than .05 but less than .15, we could say that the direction *leans* in a certain direction. If p is greater than .15 but less than .25, we could say that there is a *hint* about the true direction. Tukey was

not suggesting that we should use .25 as the level of significance. Rather, he was telling us to stop treating statistical testing as an all-or-nothing procedure and instead use appropriate wording to describe degrees of uncertainty.

Tukey’s (1969, 1991) advice incorporates a great deal about what modern psychological investigations have told us about how humans understand probability. Modern concepts of probability began with Kolmogorov’s (1933) mathematical definition of probability as a measure of sets in an abstract space of events. Although all mathematical properties of probability can be derived from this definition, it is of little value in helping us to apply probability to real-life situations. Comprehending how humans understand probability was helped enormously by the concept of “personal probability” proposed almost a half century ago by both de Finetti (1974) and Savage (1954), who contended that probability is a common concept that people can use coherently if their inferences using it follow a few simple rules. Unfortunately, in a series of ingenious experiments, the psychologists Kahneman and Tversky (summarized in Kahneman et al., 1982) found no one whose probabilistic judgments met Savage’s (1954) criteria for coherence. They found, instead, that most people did not have the ability to keep a consistent view of what different numerical probabilities meant. They reported that the best humans could manage was a vastly simplified probability model, which they attribute to Suppes, that met Kolmogorov’s axioms and fit their data. Suppes’s model has only five probabilities:

1. Surely true.
2. More probable than not.
3. As probable as not.
4. Less probable than not.
5. Surely false.

Although Suppes’s model has the benefit of fitting Kahneman et al.’s (1982) data, it also leads to a remarkably uninteresting mathematical theory with only a few possible theorems. Indeed if Suppes’s model is the only one that fits personal probability, then many of the techniques of statistical analysis that are standard practice are useless, because they serve only to produce distinctions below the level of human perception. In view of these results, Tukey’s approach to interpreting p values may in-

deed be the only sensible way to go; arguing about .04 or .05 or .06 is a poor use of time.

Finally, we mentioned earlier that a widely cited criticism of NHST is that research consumers frequently mistakenly equate statistical significance with scientific significance. Researchers can help to alleviate this problem by avoiding use of confusing terminology. For example, rather than describing results as “boys scored significantly higher than girls, $p < .05$,” authors could simply say, “on average, boys scored higher than girls, $p < .05$ ” (Robinson, Levin, Halbur, & O’Ryan, 2001).

One Expanded View of NHST

Recently, Jones and Tukey (2000), expanding on an old idea (e.g., Lehmann, 1959; Wald, 1947), suggested a better way to interpret significant and nonsignificant p values. If p is less than .05, researchers can conclude that the direction of a difference was determined: either the mean of group 1 is greater than the mean of group 2, or vice versa. If p is greater than .05, the conclusion is that the sign of the difference is not yet determined. This trinary decision approach (either $\mu_1 > \mu_2$, $\mu_2 > \mu_1$, or do not know yet) has the advantages of stressing that research is a continuing activity and never having to “accept” a null hypothesis that is likely untrue. Fisher (1929) also commented on NHST’s inability to support a null hypothesis as true:

For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning . . . it would therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data; but that they are never capable of establishing them as certainly true . . . (p. 192)

Rather than concluding that “there was no difference among the treatments ($p = .07$)” or that “the two variables were not correlated ($p = .06$),” authors should instead state that “the direction of the differences among the treatments was undetermined” or that “the sign of the correlation among the two variables was un-

determined.” This language avoids leaving the impression that the null hypothesis was accepted and suggests rather that more data are needed before a determination can be made.

Conclusions and Recommendations

NHST, as currently constituted, is a tool of limited usefulness. It is useful in determining the direction of an effect. It can be a valued accompaniment to effect sizes by providing information about the trustworthiness of estimates of the size of the effect. It is not very useful when sample sizes are extremely large. On the other hand, effect sizes are not particularly helpful when testing model fit. In addition accurate estimates of effect sizes are sometimes impossible to obtain, as for example in medical research where the continued use of a control group is not ethical.

NHST, like any other statistical procedure, is not well used with individual research studies when the goal is to predict long-term frequencies of occurrence. However, modified versions of NHST can be used to good effect, as in tests on means with a trinary hypothesis. Such procedures have been in use for decades in sequential analysis (e.g., it's better, it's worse, or keep on testing). NHST is best used in conjunction with a series of investigations. Replicated significant results serve as the foundation of scientific justification of the direction of an effect. Replications with extensions also serve to enhance the generalizability of results while adding to the evidence for the effect.

We wish to note a key difference between what we are advocating as a research model, based on Fisher's ideas, and what Thompson (2002b) recently referred to as thinking meta-analytically. Thompson prefers a model where effect sizes from individual studies are interpreted in the context of previous studies. This minimizes reliance on NHST and focuses instead on size of the effect. Our goal is for researchers to first think “programmatically” where they embark on a series of experiments and use NHST to evaluate effect direction. Because this involves fine-tuning in altering the design, treatment, sample size, etc., with each successive replication to both produce the most effective treatment and increase generalizability, the researcher is less interested in size of the effect and instead focuses on obtaining a reliable effect.

Once a treatment is shown to be reliable in the laboratory or classroom setting, it is then ready to be tested in schools and districts using randomized field trials (Boruch, de Moya, & Snyder, 2002). Once randomized field trials have been conducted, we can begin to think meta-analytically.

Last, it has been our informal experience that many users of NHST interpret the result as the probability of the null hypothesis based upon the data observed, that is, $P(H_0|\text{data})$. But what is actually yielded is the probability of the data given the null hypothesis, $P(\text{data}|H_0)$. This error suggests that users really want to make a different kind of inference—a probabilistic statement of the likelihood of the hypothesis. To be able to make such inferences requires transforming the usual $P(\text{data}|H_0)$ with a straightforward application of Bayes's theorem, which directly allows us to estimate $P(H|\text{data})$ from $P(\text{data}|H)$: $P(H|\text{data}) = \frac{P(\text{data}|H) P(H)}{P(\text{data})}$. The ratio of the unconditioned probability of the hypothesis of interest, $P(H)$, to the unconditioned probability of the data, $P(\text{data})$, is the price we have to pay to be able to make the inference we want. $P(H)$ is the prior probability of the hypothesis and $P(\text{data})$ is the prior probability of the data, and we must estimate it somehow.

The famous statistician Jimmie Savage (1962) was referring to the estimations of priors when he said that you cannot eat the Bayesian omelet without breaking the Bayesian egg. This is the second important difference. Somehow we must determine how likely we thought our hypothesis was before we gathered these data. When the data were gathered as part of a long-term investigation we can use other results to provide an estimate. When we are working fresh we must choose a prior in some other way. There is an extensive literature that details various approaches to estimating this component (Box & Tiao, 1973; Novick & Jackson, 1974; Winkler, 1993). A common approach parallels standard hypothesis testing by looking at the ratio of two exhaustive and incompatible hypotheses $\frac{P(H_1|\text{data})}{P(H_2|\text{data})} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_2)} \times \frac{P(H_1)}{P(H_2)}$. The first term on the right is the likelihood ratio and the second term is the ratio of the prior probability of the competing hypotheses (the odds ratio). Note that $P(\text{data})$ drops out and we have a somewhat easier task since we

often have strong prior beliefs about the likelihood of competing hypotheses. Although this approach may seem to require an unfamiliar mode of operation initially, the evidence that has accumulated on the use of Bayesian methods since Savage's prophetic words strongly suggests that the quality of the omelet makes it worthwhile to break the egg.

NOTES

The authors wish to thank Shlomo Sawilowsky for his intellectual generosity and helpful comments on an earlier draft of this article and also the editors and anonymous reviewers for their helpful comments that forced us to sharpen our arguments, include relevant literature that we had previously missed, and more accurately say what we meant.

¹ The assumption that the speed of light is constant in all reference frames was critical to Einstein's deductions in what became his Theory of Special Relativity.

² If the situation of concern to us matches most of what we observe in the everyday world—not being too fast, not being too small, not being too cold or too warm—Newton's Laws work very well indeed.

REFERENCES

- Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.
- Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.
- Alberoni, F. (1962). Contribution to the study of subjective probability. Part I. *Journal of General Psychology*, 66, 241–264.
- Anscombe, F. (1963). Tests of goodness of fit. *Journal of the Royal Statistical Society B*, 25, 81–94.
- Arbuthnot, J. (1710). An argument for divine providence taken from the constant regularity in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27, 186–190.
- Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized field trials in education research*. Washington, DC: Brookings Institution.
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.
- Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society B*, 135, 1–45.

- Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS: 88. *Chance*, 14(1), 10–21.
- Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287–292.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics*, 4, 49–63.
- de Finetti, B. (1974). *Theory of probability*. (A. Machi & A. Smith, Trans.) New York: Wiley. (Original work published 1970)
- Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal*, 37, 747–773.
- Fidler, F. (2002). The fifth edition of the *APA Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, 62, 749–770.
- Fisher, R. A. (1970). Theory of statistical estimation. *Proceedings of the Cambridge Philosophic Society*, 22, 700–725. (Original work published in 1925)
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.
- Fisher, R. A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research*, 39, 189–192.
- Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1, 379–390.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Graunt, J. (1662). *Natural and political observations on the bills of mortality*. London: John Martyn and James Allestry.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology*, Vol. 5 (pp. 1–22). Stamford, CT: JAI Press.
- Johnston, F. R. (2000). Word learning in predictable text. *Journal of Educational Psychology*, 92, 248–255.
- Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5, 411–414.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgments under uncertainty: heuristics and biases*. Cambridge, England: Cambridge University Press.
- Kaiser, H. (1970). A second generation little jiffy. *Psychometrika*, 35, 411–436.
- Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, 70, 65–79.
- Kolata, G. (2001, November 17). Constants of the universe may be changing. *The New York Times*, p. A1.
- Kolmogorov, A. N. (1933). *Grundbegriffe der wahrscheinlichkeitsrechnung*. Berlin, Germany: Springer-Verlag.
- Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, 43, 635–642.
- Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York: Wiley.
- Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: the elimination of Type IV errors. *American Educational Research Journal*, 7, 397–421.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. Chicago: Aldine.
- Ni, Y. (2001). Semantic domains of rational numbers and the acquisition of fraction equivalence. *Contemporary Educational Psychology*, 26, 400–417.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- Novick, M. R., & Jackson, J. E. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.
- Orange, C. (1999). Using peer modeling to teach self-regulation. *Journal of Experimental Education*, 68, 21–39.
- Research in the Schools* (1998). 5(2), complete issue.
- Robinson, D. H., Fouladi, R. T., Williams, N. J., & Bera, S. J. (2002). Some effects of providing effect size and “what if” information. *Journal of Experimental Education*, 70, 365–382.
- Robinson, D. H., Funk, D. C., Halbur, D., & O’Ryan, L. (in press). The .05 level of significance in educational research: Traditional, arbitrary, sacred, magical, or simply psychological? *Research in the Schools*.
- Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 26(5), 21–26.
- Robinson, D. H., Levin, J. R., Halbur, D., & O’Ryan, L. (2001). Does use of statistical language constitute a “significant” roadblock to readers’ interpretations of research results? *Journal of Educational Psychology*, 93, 646–654.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Thousand Oaks, CA: Sage.
- Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice*, 2(2), 133–150.
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166–169.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–1284.
- Royer, J. M. (2000). Editorial: A policy on reporting of effect sizes. *Contemporary Educational Psychology*, 25, 239.
- Rubman, C. N., & Waters, H. S. (2000). A, B seeing: The role of constructive processes in children’s comprehension monitoring. *Journal of Educational Psychology*, 92, 503–514.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Savage, L. J. (1962). “Subjective probability and statistical practice.” *The foundations of statistical inference: A discussion*. London: Methuen and Co., and New York: John Wiley and Sons, Inc., 9–35.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115–129.
- Schneider, W., Roth, E., & Ennemoser, M. (2000). Training phonological skills and letter knowledge in children at risk for dyslexia: A comparison of three kindergarten intervention programs. *Journal of Educational Psychology*, 92, 284–295.
- Selvin, H. C. (1957). A critique of tests of significance in survey research. *American Sociological Review*, 22, 519–527.
- Shaver, J. (1985). Chance and nonsense: A conversation about interpreting tests of statistical significance, part 2. *Phi Delta Kappan*, 67(2), 138–141.
- Slavin, R. E., & Madden, N. A. (Eds.). (2001). *One million children: Success for All*. Thousand Oaks, CA: Corwin.
- Sohn, D. (2000). Significance testing and the science. *American Psychologist*, 55, 964–965.
- Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians’ health study. *New England Journal of Medicine*, 318, 262–264.
- Student. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 29, 363–379.
- Thompson, B. (1994). *The concept of statistical significance testing* (Report No. EDO-TM-94-1). Washington, DC: Office of Educational Research and Improvement. (Grant No. RR93002002.) (ERIC Document Reproduction Service No. 366 654)

- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25(2), 26–30.
- Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher*, 26(5), 29–32.
- Thompson, B. (2002a). “Statistical,” “practical,” and “clinical”: How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64–71.
- Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 24–31.
- Thompson, B. (2003). Various editorial policies regarding statistical significance tests and effect sizes. Retrieved March 20, 2003, from <http://www.coe.tamu.edu/~bthompson>
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24, 83–91.
- Tukey, J. W. (1991). The philosophy of multiple comparisons, *Statistical Science*, 6(1), 98–116.
- Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, 4, 212–213.
- Wald, A. (1947). *Sequential analysis*. New York: Wiley.
- Wang, A. Y., & Newlin, M. H. (2000). Characteristics of students who enroll and succeed in psychology web-based classes. *Journal of Educational Psychology*, 92, 137–143.
- Webb, J. K., Murphy, M. T., Flambaum, V. V., Dzuba, V. A., Barrow, J. D., Churchill, C. W., et al. (2001). Further evidence for cosmological evolution of the fine structure constant. *Physical Review Letters*, 87, 91, 301.
- Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Winkler, R. L. (1993). Bayesian statistics: An overview. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 201–232). Hillsdale, NJ: Erlbaum.
- Zabell, S., & Wainer, H. (2002). A small hurrah for the Black Death. *Chance*, 15(4), 58–60.

AUTHORS

HOWARD WAINER is Distinguished Research Scientist, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; hwainer@nbme.org. His interests include psychometrics, graphics, and statistics.

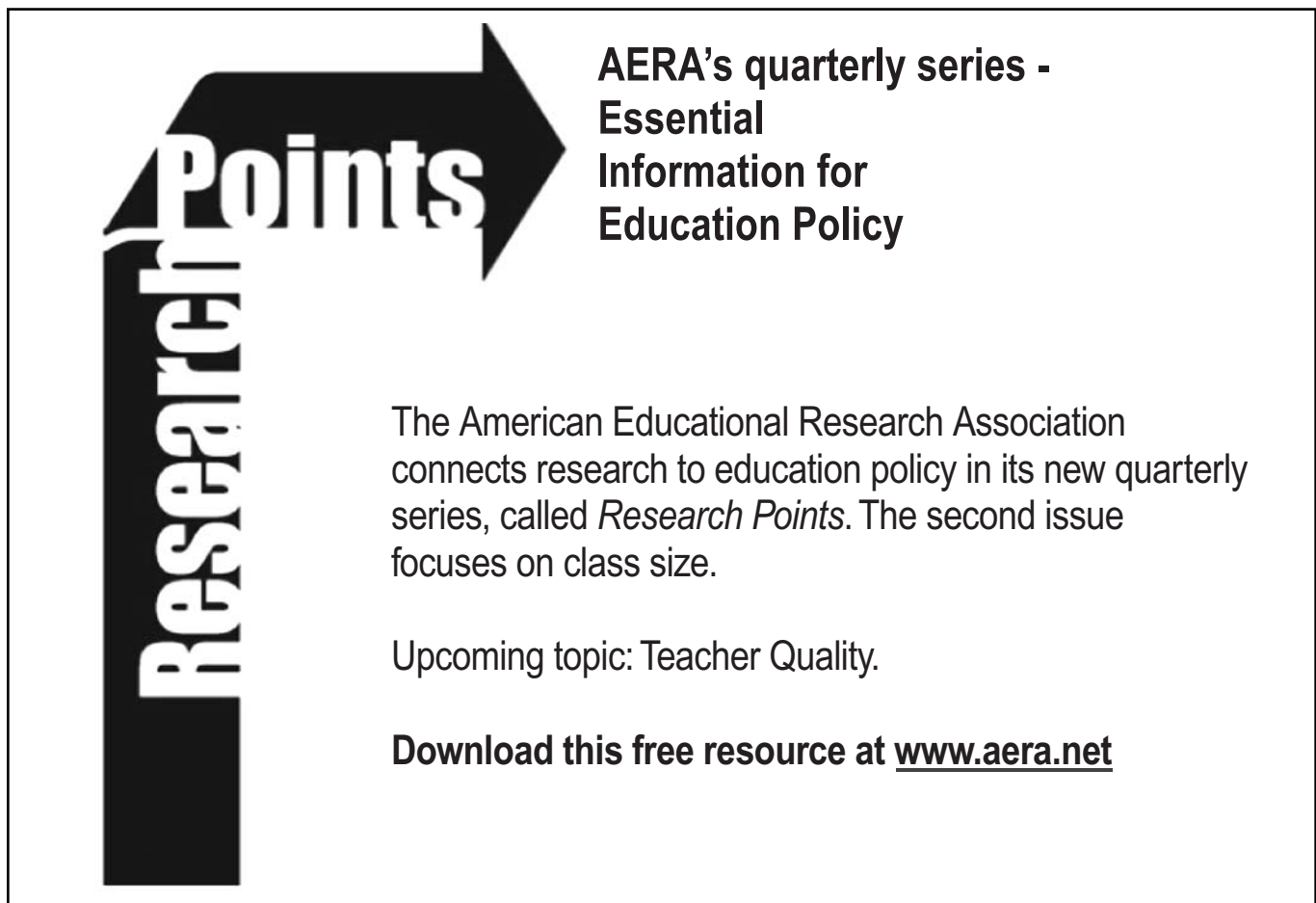
DANIEL H. ROBINSON is an associate professor of educational psychology at the University of Texas, Austin, TX 78712; dan.robinson@mail.utexas.edu. His research interests include web-based text comprehension strategies and the communication of research results.

Manuscript received September 9, 2002

Revisions received January 1, 2003;

June 1, 2003

Accepted June 19, 2003



Research Points

**AERA's quarterly series -
Essential
Information for
Education Policy**

The American Educational Research Association connects research to education policy in its new quarterly series, called *Research Points*. The second issue focuses on class size.

Upcoming topic: Teacher Quality.

Download this free resource at www.aera.net